PROCEDURE FOR CHARACTERIZING A SOUND SIGNAL

The invention relates to a method for characterizing, according to specific parameters, a sound signal developing over time in different frequency bands.

The field of the invention is that of sound signal recognition applied in particular to the identification of musical works used without authorization.

In fact, the development of methods of digitizing and multimedia have caused a considerable increase in such fraudulent uses. The result is a new problem for those agencies charged with collecting royalties, since there must be some way to identify these uses, especially on the interactive digital networks such as the Internet, in order to satisfactorily assess and to distribute the compensation due to the authors of these musical works.

Consequently, in order not to be limited to musical works, a sound signal is more generally considered.

The object of the present invention is then to create a database of sound signals, each sound signal being characterized by one fingerprint such that being given a unknown sound signal that is characterized in this same fashion, a search can be executed and a rapid comparison of the fingerprint of said

unknown signal made with the universe of fingerprints in the database.

The fingerprint is constituted of specific parameters determined in the following fashion.

In a first step, the sound signal is broken down in that its amplitude x(t) varies with time t, according to different frequency bands k : x(k, t) is the amplitude of the sound signal filtered into the frequency band k and represented in Fig. 1a.

As represented in Fig. 1c, the short-term energy E(k, t) of this filtered sound signal is calculated using a window h(t) represented in Fig. 1b, having a support of 2N seconds. This calculation is repeated by sliding said window every S seconds.

These values E(k, t) constitute the specific parameters of an extract of 2N seconds of the sound signal x(k, t) in the frequency band k.

Other parameters can be obtained by calculating the energy of E(k, t) for the different frequency bands j by using a window h'(t) represented in Fig. 2b, having a base of 2N' seconds; this calculation is reiterated by sliding said window every S' seconds : one then obtains F(j, k, t) represented in Fig. 2. These F(j, k, t) values are standardized with respect to their maximum in order to make them independent of the amplitude of the sound signal.

Thus standardized, these values constitute specific parameters of an extract of 2N' seconds of the sound signal x(k, t) in the k band of frequencies.

One can also calculate the phase of E(k, t) for different bands of frequencies j : one obtains P(j, k, t). The P(j, k, t) values are standardized with respect to a reference value P(1, j, t) and one then obtains other specific parameters of an extract of 2N' seconds of sound signal.

Other parameters can be added such as the mean value of the E(k, t) energy.

The object of the invention is a method for characterizing in accordance with specific parameters a sound signal x(t) evolving according to the time t over a duration D in different bands of frequencies k and then written x(k, t), principally characterized in that it consists of storing the signal x(t), calculating the energy E(k, t) of said signal x(k, t) for each of said bands of frequencies k, k varying from 1 to K and according to a temporal window h(t) of a duration of 2N, storing the values of the energy E(k, t) obtained, these values constituting the specific parameters of an extract of a duration of 2N of the sound signal x(t) and reiterating this calculation at regular intervals, in order to obtain the universe of specific parameters for the duration D of the sound signal x(t).

In addition, it consists of calculating and storing the energy F(k, j, t) of E(k, t) for the bands of frequencies j, j varying from 1 to J, according to a temporal window h'(t) of a duration of 2N', the JxK values of the energy F(j, k, t) obtained constitute the specific parameters of an extract of a duration of 2N' of the sound signal x(t) and reiterating this calculation at regular intervals, in order to obtain the universe of specific parameters for the duration D of the sound signal x(t).

It may consist of calculating the phase P(j, k, t) of the energy E(k, t) for the bands of frequencies j, j varying from 1 to J with j being different from k, and including the values of the phase P(j, k, t) obtained among the specific parameters of the sound signal x(t).

It can also consist of calculating the mean value of the energy E(k, t) over 2N' seconds for each frequency band j, in reiterating this calculation at regular intervals, in order to obtain the universe of specific parameters for the duration D of the sound signal x(t) and including the mean values so obtained among the specific parameters of the sound signal x(t).

According to one feature, it consists of taking into account the specific parameters of a sound signal x(t) as the components of a vector representing x(t), of positioning the vectors in a space of as many dimensions as there are parameters, of defining classes including the most proximate vectors and of recording said classes.

The classes having inter-class distances and intra-class distances, the method consists advantageously of selecting from among the specific parameters those parameters making it possible to obtain the relatively large inter-class distances with respect to of the intra class distances and of recording the selected parameters.

The invention relates also to a device for identifying a sound signal, characterized in that it comprises a database service comprising means for implementing the method for characterizing a sound signal according to specific parameters

as described hereinbefore and the means for executing a search for said signal in the database.

Preferably, the search means comprise means for directly recognizing the class to which said sound signal belongs and means for executing a search for the class by comparison of the specific parameters of the unknown sound signal with those of the database, the class being chosen, for example, using the method of the nearest neighbor algorithm.

Other characteristics and advantages of the invention will become more apparent when reading the description provided by way of example and non-limitingly and with reference to the appended drawings, wherein :

Figs. 1a, 1b and 1c represent, respectively, the diagrammatic plottings of the variation of a sound signal $x(k_i, t)$ filtered into a band of frequencies k_i, a Hamming window $h(t)$ and the short-term energy $E(k_i, t)$ of the signal $x(k_i, t)$;

Figs. 2a, 2b and 2c represent, respectively the diagrammatic plottings of the variation of energy $E(k_i, t)$ for the frequency band k_i, a Hamming windos $h'(t)$ and the energy $F(j_m, k_i, t)$ of $E(k_i, t)$ for the band of frequencies j_m.

Fig. 3 diagrammatically represents the universe of vectors $V[x(t)]$ constituting the fingerprint of a signal $x(k, t)$;

Fig. 4 diagrammatically represents the storing of fingerprints;

Fig. 5 represents the classification of the sound signals according to two parameters;

Fig. 6 represents a method for searching for a sound signal using the method of the nearest neighbor algorithm;

Fig. 7 diagrammatically represents a database service for storing the fingerprints of the sound signals.

The sound signals that are processed according to this method of characterization are recorded sound signals, particularly on compact disks.

In the following, it will be considered that the sound signal x(t) is a digital signal sampled at a sampling frequency of f_e, for example 11,025 Hz corresponding to one quarter of the current sampling frequency for compact disks, which is 44,100 Hz.

Therefore, an analog sound signal can be characterized : it must first be converted into a digital signal by means of an analog - digital converter.

The sound signal x(k, t) represented in Fig. 1a for $k = k_i$ is thus a digital signal sampled at the frequency fe and obtained after filtering into a band of frequencies k_i. Each value of this digital signal sampled is coded, for example, in 16 bits. The bands of frequencies are bands of the audible spectrum varying from approximately 20 Hz to 20 kHz and sectioned into K (k varies from 1 to K) bands of frequencies, K = 127, for example.

The short-term energy E(k, t) represented in Fig. 1c for $k = k_i$ is calculated using a window h(t) of 2N seconds; for

example, a Hamming window having a base of approximately 23 ms represented in Fig. 1b.

E(k, t) is the square of the module of a transformation of the sound signal sampled x(t) in the time - frequency plan or in the time - scale plan. Among the transformations that can be utilized are the Fourier transformation, the cosine transformation, the Hartley transformation and the wavelet transformation. A bank of band-pass filters also does this type of transformation. The short-term Fourier transformation makes possible a time - frequency representation adapted to the musical signal analysis. Accordingly, the energy E(k, t) is written :

$$E(k,t) = \left| \sum_{n=-N}^{n=N} x(t + n/f_e) \cdot h(n/f_e) \cdot e^{-4i\pi kn/N} \right|^2$$

wherein i such that $i^2 = -1$

One slides the window over the sound signal every S seconds; for example, every 10 ms. E(k, t) will thus be sampled every 10 ms : $E(k, t_0)$, $E(k, t_1)$ with $t_1 = t_0 + 10$ ms, etc. will be obtained.

Thus, all of the S seconds of the sound signal x(t) will be coded by a vector having K components E(k, t), each of these components coding for the energy of 23 ms or the sound signal x(t) in K bands of frequencies.

Other parameters are obtained by reproducing in any fashion the aforementioned calculations and applying them each time to E(k, t) as represented n Figs. 2a to 2c.

The energy E(k, t) is filtered into J different bands of frequencies : E(j, k, t) is the energy E(k, t) filtered into the band of frequencies j, j varying from 1 to J with, for example, J = 51.

Then F(j, k, t) is calculated, represented in Fig. 2c), for $k = k_i$ and $j = j_m$, using a window h'(t) of 2N' seconds; for example a Hamming window having a base of 10 s. Thus, using i such that $i^2 = -1$, one can write :

$$F(j,k,t) = \left| \sum_{n=-N'}^{n=N'} E(k,t+n/f_{e'}) \cdot h'(n/f_{e'}) \cdot e^{-4i\pi jn/N'} \right|^2$$

In our example, every seconds (S' = 1), the sound signal x(t) is coded by 127x51 parameters F(j, k, t), each real F(j, k, t) representing the energy of ten seconds (2N' = 10) of the energy signal E(k, t) in the frequency band j.

In order to make F(j, k, t) independent of the amplitude of the signal that can be more or less strong, these values are put in relation to a reference value; in the present case, the maximum value of $F_M(j, k, t)$ for all of the k and j taken into account. Thus KxJ parameters are obtained $F(j, k, t)/F_M(j, k, t)$.

In addition, the phase of the energy E(k, t) in each of the bands of frequency j is calculated every 2N' seconds : P(j, k, t).

To do this, the argument of the Fourier transformation of E(k, t) in each of the frequency bands j is calculated :

$$P(j,k,t) = Arg \left| \sum_{n=-N'}^{n=N'} E(k,t+n/f_{e'}) \cdot h'(n/f_{e'}) \cdot e^{-4i\pi jn/N'} \right|$$

As above, these values are put in relation to a reference value; in the present case, the value of P(j, k, t) for the second band of frequencies (j = 1) considered, because the temporal reference of the sample is unknown : the origin of the time is unknown.

To do this, the phases yielded φ(j, k, t) are calculated using the following formulae :

$$\varphi(1, k, t) = P(1, k, t)$$

$$\varphi(j, k, t) = P(j, k, t) - P(1, k, t) \cdot \frac{f(k)}{f(1)}, \text{ for } k > 1$$

wherein the f(k) are the central frequencies of channels k.

Thus, KxJ parameters corresponding to the values of the yielded phase φ(j, k, t) are obtained.

Other parameters can also be taken into account; in particular, the mean values of the energy E(k, t) over 2N' seconds and this for each band of frequencies j : E(j, k, t).

The universe of these standardized parameters define at regular intervals a fingerprint that can be considered as a vector V(x(t)). The universe of the standardized parameters, for example, $F(j, k, t)/F_M$ and P(j, k, t)-P(j, 1, t) define every S' seconds a fingerprint that can be considered as a vector V(x(t)) having 2xKxJ dimensions (2x127x51) or about 13,000 in our example), one dimension per parameter, each vector characterizing an extract of 2N'seconds of the sound signal x(t), 10 seconds in our example.

This characterization is reiterated every S' seconds, every second for example (S' = 1).

As represented in Fig. 3, a signal x(t) over T seconds is ultimately characterized by L vectors V, L being approximately equal to T/S'.

For a sound signal lasting 10 mn or 600 s, 600 vectors are obtained; that is, 600x2xJxK parameters.

These vectors are stored in the storage zone 10 of a database housed on a server or on a compact disk. Fig. 4 represents the universe of the vectors V of a signal or of a work A by VA, likewise VB for a work B, etc.

It is desirable to reduce the number of components of these vectors; in other words, the number of parameters in order to obtain a vector or a fingerprint of a more reduced size in view of its storage in the database. Furthermore, when it is a question of comparing the fingerprint of an unknown sound signal to those in the database, it is desirable that the number of parameters to be compared be reduced in order that said search can be executed quickly.

Now, these parameters do not all contain the same quantity of information; certain ones can be redundant or useless. That is why one selects the most meaningful parameters from among all parameters, using a mutual information calculation presented in the publication PROC. ICASSP '99, Phoenix, Arizona, USA, March 1999 H.YANG, S. VAN VUUREN, H. HERMANSKY, "Relevancy of Time - Frequency Features for Phonetic Classification Measured by Mutual Information". Thus, K to K_1 and J to J_1 are limited.

A method for selecting these parameters will now be presented.

Each of the fingerprints of these sound signals; that is, each of these vectors is classified into a space R to N dimensions, N being the number of components of the vectors. For the sake of simplicity, an example of classification for vectors having 2 dimensions P1 and P2 is represented in Fig. 5.

The classes C(m) are defined by grouping the vectors by proximity, m varying from 1 to M. For example, one can decide that one class corresponds to one musical work : in this case M is the number of musical works stored in the database.

The result of the mutual information calculation between these classes C(m) and the parameters is that the relevance of the parameters is linked to the inter and intra class distances : relevant parameters assuring relatively large inter-class distances d compared to the intra-class distances D.

By keeping only the relevant parameters, K_1 and J_1 are thus defined.

For example, one can consider five ($K_1 = 5$) bands of frequencies centered on 344 Hz, 430 Hz, 516 Hz, 608 Hz and 689 Hz, respectively.

Tests have been done by taking $J_1 = 3$.

The classes C(m) are thus constituted using the vectors $V_q(x)$ not comprising more than $2 \times K_1 \times J_1$ components.

An example will be given for $K_1 = 5$ and $J_1 = 3$, of the size of the memory of a database containing 1,000 hours of music and by taking into account as parameters E(k, t) and F(j, k, t), each of these parameters being coded using 4 bytes.

The E(k, t) parameters calculated every 10 ms occupy 1,000 x 3,600 x 100 x 4 bytes or apprximately 7 gigabytes.

The parameters F(j, k, t) calculated every second occupy 1,000 x 3,600 x 3 x 5 x 4 bytes or approximately 200 megabytes.

These parameters are associated with sound signal references : if one considers that the references contain 100 characters each coded on one byte, these references occupy 1,000 x 10 x 100 bytes or approximately 1 megabyte.

Such a database would ultimately occupy approximately 7 gigabytes.

When one wishes to identify an unknown sound signal, one first of all establishes the fingerprint, referenced V(xinc) in Fig. 6, as described hereinbefore, knowing that the unknown sound signal can be a complete musical work or an extract therefrom.

The search for the class of this fingerprint in the database thus consists, according to a classical method illustrated in Fig. 6, of comparing the parameters of this fingerprint V(xinc) to those of the fingerprints of the database. The most proximate fingerprints, called the nearest neighbors, define the class in the following fashion : the class is that of the majority of the nearest neighbors.

A database server 1 is diagrammatically represented in Fig. 7. It comprises a storage zone 10 for the data of the database, in which the fingerprints of the mixed sound signals are stored by their references. In addition, it comprises a memory 11, the aforementioned characterization and programs are stored, a processor 12 with working memories for deploying the programs. It obviously comprises an I/O interface 13 and a bus 14 connecting these diverse elements with each other.

When entering new sound signals into the database 1, the interface 13 receives the signal x(t) accompanied by its references; if it is only an unknown signal to be identified, the interface 12 receives only the unknown signal x(t).

Upon output, the interface 13 provides a response to the search for an unknown signal. This response is negative if the unknown signal does not exist in the storage zone 10; if the signal has been identified, the response includes the references of the identified signal.